KnowIT VQA: Answering Knowledge-Based Questions about Videos

Noa Garcia — Osaka University
Mayu Otani — CyberAgent, Inc.
Chenhui Chu — Osaka University
Yuta Nakashima — Osaka University

https://knowit-vqa.github.io/

Motivation

Current VQA limitations:

1) Temporal coherence → **VideoQA**

2) External Knowledge → **KBVQA**

So far, they are independent problems.

We address VideoQA and KBVQA **together**.

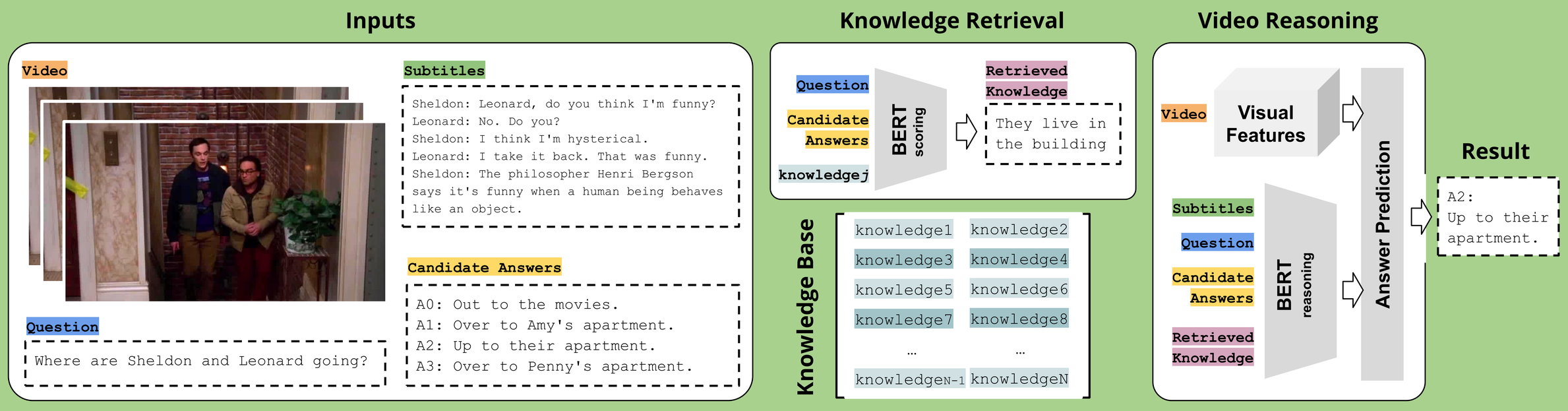
KnowIT VQA Dataset

- Human expert annotators.
- 207 episodes, 9 seasons.
- 12,087 video clips.
- 24,282 questions pairs.

For each question:

- 4 candidate answers.
- Grounded knowledge.
- Knowledge type (recurrent, specific)

ROCK Model



BERT scoring input

$$[CLS] + q_i + a_i^0 + a_i^1 + a_i^2 + a_i^3 + [SEP] + kg_j + [SEP]$$

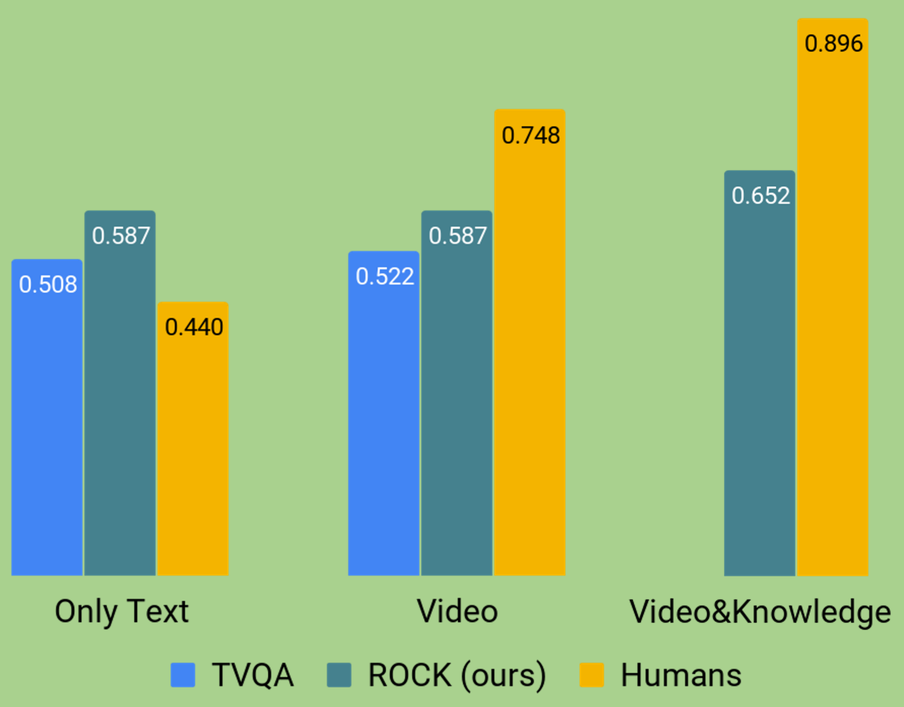
BERT reasoning input

$$[CLS]\ (+\ caps) + subs + q + [SEP] + a^c + kg + [SEP]$$

Results and Examples

Visual features:

- *ResNet,*
- *Facial,*
- *Concepts,*
- *Captions*.



Penny: What are you doing at work these days?
Sheldon: Oh. I'm working on time-dependent backgrounds in string theory. Specifically, quantum field theory in D-dimensional de Sitter space.

What night is it?

Wednesday

Monday

Friday

Saturday ✓

Retrieved Knowledge

Saturday is Sheldon's laundry night.